

PATENT APPLICATION

A FIBRE CHANNEL SWITCH THAT ENABLES END DEVICES IN DIFFERENT FABRICS TO COMMUNICATE WITH ONE ANOTHER WHILE RETAINING THEIR UNIQUE FIBRE CHANNEL DOMAIN_IDS

Dinesh G. Dutt
1176 Corral Ave.
Sunnyvale, CA 94086
Citizenship: India

Silvano Gai
3021 Mauna Loa Ct.
San Jose, CA 95132
Citizenship: Italy

Bruno Raimondo
655 Lynwood Avenue #2
Mountain View, CA
94043
Citizenship: Italy

Thomas Edsall
13208 Peacock Ct.
Cupertino, CA 95014
Citizenship: United States

Subrata Banerjee
2808 Morgan Drive
San Ramon, CA 94583
Citizenship: United States

Rajeev Bharadwaj
479 Senna Court
Sunnyvale, Ca 94086
Citizenship: United States

Ankur Jain
1308 Fairway Entrance Dr.
San Jose, CA 95131
Citizenship: India

Davide Bergamasco
825 E. Dana Street
Mountain View, CA 94041
Citizenship: Italy

Assignee:
Andiamo Systems, Inc.
375 East Tasman Drive
San Jose, CA 95134

A Delaware corporation

Status: Large Entity

A FIBRE CHANNEL SWITCH THAT ENABLES END DEVICES IN DIFFERENT FABRICS TO COMMUNICATE WITH ONE ANOTHER WHILE RETAINING THEIR UNIQUE FIBRE CHANNEL DOMAIN_IDs

Related Applications

[0001] The present invention is a Continuation-in-Part of co-pending, commonly assigned, application Serial Number 10/034,160 filed December 26, 2001 and entitled "Methods and Apparatus for Encapsulating a Frame for Transmission in a Storage Area Network", incorporated herein for all purposes.

BACKGROUND OF THE INVENTION

Field of the Invention

[0002] The present invention relates to storage area networks, and more particularly, to a Fibre Channel Switch that enables the end devices in different Fabrics to communicate with one another while retaining their unique Fibre Channel IDs.

Background of the Invention

[0003] With the increasing popularity of Internet commerce and network centric computing, businesses and other organizations are becoming more and more reliant on information. To handle all of this data, storage area networks or SANs have become very popular. A SAN typically includes a number of storage devices, a plurality of Hosts, and a number of Switches arranged in a Switching Fabric that connects the storage devices and the Hosts.

[0004] Most SANs rely on the Fibre Channel protocol for communication within the Fabric. For a detailed explanation of the Fibre Channel protocol and Fibre Channel Switching Fabrics and Services, see the Fibre Channel Framing and Signaling Standard, Rev 1.70, American National Standard of Accredited Standards Committee (NCITS), February 8, 2002, and the Fibre Channel Switch Fabric - 2, Rev. 5.4,

NCITS, June 26, 2001, and the Fibre Channel Generic Services – 3, Rev. 7.01, NCITS, November 28, 2000, all incorporated by reference herein for all purposes.

[0005] In Fibre Channel, each device (Hosts, storage devices and Switches) is identified by a globally unique, eight (8) byte wide World Wide Name (WWN) assigned by the manufacturer. There are two kinds of WWNs used in FC networks. If you consider a device with one or more FC adapters (or HBAs or ports) to connect to a FC network, every device is assigned a node WWN (nWWN) and each adapter is assigned a port WWN (pWWN). The nWWN and pWWN are different from each other. When the Fibre Channel devices are interconnected to form a SAN, the WWN (along with other parameters) is the primary mechanism to identify each device. Fibre Channel frames are used for communication among the devices in the SAN. The WWN, however, is not used by the frames. Each adapter or port must login to the FC network. At this time, each port is dynamically assigned a unique Fibre Channel address (FC_ID) by the Fabric. The FC_ID is used in FC networks for end devices to address each other.

[0006] The three byte wide Fibre Channel addresses are hierarchically structured in three fields, each one byte long: Domain_ID, Area_ID, and Port_ID. Each Switch within the Fabric is assigned a Domain_ID. The end devices attached to a particular Switch are assigned the Domain_ID of that Switch. The Switch manages the allocation of the Area_ID and Port_ID fields for each end device to guarantee the uniqueness of the assigned addresses in that Domain. For example, if a Switch is assigned a Domain number five and the Switch subdivides its address space in two areas each having three connected end devices, then a possible Fibre Channel address allocation is: 5:1:1, 5:1:2, 5:1:3, 5:2:1, 5:2:2, and 5:2:3.

[0007] Fibre Channel based SANs are often organized into zones. Within each zone, Hosts can see and access only storage devices or other Hosts belonging to that zone. This allows the coexistence on the same SAN of different computing environments. For example, it is possible to define on a SAN a Unix zone and a separate Windows zone. Unix servers belonging to the Unix zone may access only storage or Hosts devices within the Unix zone, and do not interfere with the other devices connected to the SAN. In the same manner, Windows servers belonging to the Windows zone may access storage or Hosts devices only within the Windows zone, without interfering

with the other devices connected to the SAN. The SAN administrator may define in a SAN multiple zones, as required or dictated by the computing and storage resources connected to it. The Switching Fabric allows communications only between devices belonging to the same zone, preventing a device of one zone from seeing or accessing a device of another zone.

[0008] The information infrastructure within a large enterprise will typically have a number of SANs, each dedicated for a different organization or application within the enterprise. For example, a large corporation may have different SANs for corporate, for the sales department, the marketing department, etc. Each SAN will typically include redundant Fibre Channel fabrics connecting a plurality of Hosts and storage devices. The redundant Switches in the Fibre Channel fabrics are provided in the event a Switch or link in one Fabric goes down. If this were to occur, the redundant fabric would be used enabling normal operation of SAN. Another example is the use of a dedicated SAN for managing a mail server such as Microsoft Exchange.

[0009] The aforementioned arrangement has a number of disadvantages. Foremost, the Hosts in a given SAN can communicate only with the storage devices in that same SAN. There is no way that a Host in one SAN can directly communicate with a storage device in a second SAN. This arrangement is not only inefficient, it is expensive. Since storage devices cannot be shared among SANs, separate storage devices are required for each SAN.

[0010] The above-identified parent application partially addresses this problem by introducing the concept of a Virtual SAN or “VSAN”. The implementation of a VSAN is based on the concept of dividing the switching fabric of a single physical SAN into logical SANs, each called a VSAN. The properties of each VSAN are similar to a standard SAN, in particular: (i) unicast, broadcast and multicast traffic is confined to a VSAN and does not span multiple VSANs; (ii) Fibre Channel identifiers (FC_IDs) are assigned per VSAN. This means that a given FC address may be assigned to two different Hosts in two different VSANs; and (iii) routing and distributed Fabric Services, such as Name Server, Zone Server, etc. are maintained independently for each VSAN. This results in constraining the effect of a configuration or topology change to only the affected VSAN. Within each VSAN, a frame is forwarded as in any normal SAN, using the FC_ID.

[0011] One known solution for enabling end devices in different VSANs to communicate with one another involves the virtualization of the end devices so that there are “local instances” of each end device in the Fabric within each VSAN. See for example US Patent Publication 2003/0012204. One problem with this approach is that the border Switches between the VSANs perform FC_ID translations (i.e., Network Address translations or NATs) for the source and destination end devices. If a border Switch goes down, an alternative or fail-over path needs to be created. In addition, with certain frames, both the source and/or destination FC_IDs may be defined in the payload. A mechanism that identifies and translates these IDs must therefore be provided. This solution also does not work if encryption or a proprietary protocol is used between the source and destination end devices because there is no way for the border Switches to process the proprietary payloads or de-crypt the frames to identify the source and destination FC_IDs.

[0012] A Fibre Channel Switch and Fabric is needed which enables end devices in different Fabrics to communicate with one another while retaining their unique Fibre Channel Domain_IDs.

SUMMARY OF THE INVENTION

[0013] To achieve the foregoing, and in accordance with the purpose of the present invention, a Switch is disclosed which enables end devices in different Fabrics to communicate with one another while retaining their unique Fibre Channel Domain_IDs. The Switch is coupled to a first fabric having a first set of end devices and a second fabric having a second set of end devices. The Switch is configured to enable communication by the first set of end devices associated with the first fabric with the second set of end devices associated with the second fabric using the unique Domain_IDs of each of the first set and the second set of end devices. In one embodiment of the invention, the first and second fabrics are first and second Virtual Storage Area Networks (VSANs) respectively. In an alternative embodiment, the first fabric and the second fabric are separate physical fabrics.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] Figure 1 is an exemplary Fibre Channel Fabric according to the present invention.

[0015] Figure 2 is an exemplary VSAN topology according to the present invention.

[0016] Figure 3 is another VSAN topology illustrating VSAN border switches.

[0017] Figure 4A and 4B is an exemplary Fibre Channel frame modified for Inter-VSAN routing according to the present invention.

[0018] Figure 5 is an exemplary SAN defining “Inter-VSAN zones” according to the present invention.

[0019] Figure 6 is an exemplary VSAN for illustrating the modifications to the FSPF protocol according to the present invention.

[0020] Figure 7 is an exemplary VSAN for illustrating the operation of the present invention.

[0021] Figure 8 is a block diagram of a Switch used to implement Inter-VSAN routing according to the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0022] In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be obvious, however, to one skilled in the art, that the present invention may be practiced without some or all of these specific details. In other instances, well known process steps have not been described in detail in order not to unnecessarily obscure the present invention.

[0023] In a Fibre Channel SAN, the main services provided by the fabric include: Dynamic Address Allocation, Routing, Name Service Zone Service, and event notification. The present invention relates to Inter-VSAN and/or Inter-Fabric routing using unique Domain_IDs. Terminology specific to the present invention and defined herein includes:

Inter-VSAN Zone – a zone that contains members from multiple VSANs.

VSAN Border Switch – a switch that routes traffic to and from a pair of VSANs.

Non-adjacent VSANs – two VSANs are considered non-adjacent on a link if the link does not carry traffic between the two VSANs.

Transit VSAN – a VSAN used to carry traffic across a link between two or more non-adjacent Edge VSANs. Transit VSAN(s) thus straddle non-adjacent VSANs and enables them to communicate with one another. With this arrangement, a physical SAN is divided into lower level VSANs and high level Transit VSAN(s) for handling switching between the VSANs.

Edge VSAN – is a VSAN from which traffic is switched to or from a Transit VSAN. An Edge VSAN

can be a Transit VSAN.

[0024] Referring to Figure 1, a simplified Storage Area Network (SAN) 100 according to the present invention is shown. The SAN 100 includes a first VSAN 102 and a second VSAN 104. The first VSAN 102 includes a Switch E2 and an end device pWWN1 with a FC_ID address of 1.1.1. The second VSAN 104 includes a Switch E4 and an end device pWWN4 with a FC_ID address of 2.1.1. A border Switch B1 is also included in each VSAN. VSAN 102 and VSAN 104 are shown with only one Switch and one end device each for the purposes of illustration. It should be understood that the SAN 100 can be divided into many more VSANs, with each VSAN including one or a plurality of Switches and a set of end devices (hosts and/or storage devices where the number in the set may range from one to a multiplicity of end devices.

[0025] With the present invention, the end devices pWWN1 and pWWN4 can communicate with one another while retaining their respective FC_ID addresses. This is accomplished through the Border Switch B1 that straddles both VSANs. In a single step process, packets originating from end device pWWN1 are injected from VSAN 102 to VSAN 104 and pWWN4, and vice-versa, through the border Switch B1.

[0026] It should be noted that as used herein, the term “Fabric” generally implies a single physical Fabric that is divided into separate Virtual SANs. The two VSANs 102 and 104 as illustrated in Figure 1, could also be two physically separate Fabrics or SANs. In either case, the Border Switch enables the end devices pWWN1 and pWWN4 to communicate with one another while retaining their respective FC_ID addresses. For the remainder of the present application, the invention is described in relation to VSANs of a single Fabric or SAN. It should be understood, however, that the present invention as taught herein can be used with two separate Fabrics or SANs.

[0027] The Inter-VSAN routing using unique Domain_IDs of the present invention initially requires an administrator to define one or more Inter-VSANs and the end devices that can communicate with one another in the Fabric. After the Fabric is configured, the Border Switches: (i) exchange the Inter-VSAN routing protocol (IVRP) messages with the other Border Switches in the fabric in order to determine the topology and shortest paths to the different VSANs. In the absence of a routing

protocol, a static configuration is needed to decide the set of VSANs that must be transited by frames from an origination Edge VSAN to a terminating Edge VSAN; (ii) exchange Fabric Shortest Path First (FSPF) information between the neighbor Switches in each VSAN and the neighbor VSANs. Specifically, Border Switches inject routes into either Transit VSANs; and/or Edge VSANs connected to the Border Switch; (iii) propagate FSPF updates between the Edge and Transit VSANs only if the updates affect routes and link cost to any of the exported Domain_IDs; (iv) exchange zoning information with its neighbor Switches on linkup; (v) exchange name server databases with its neighbor Switches. Only those entries in the Inter-VSAN zone relevant to a VSAN are exchanged across VSANs; (vi) proxy as the name server for each switch in a remote Edge VSAN for queries received from Switches in the local VSAN; (vii) translates the VSAN of a frame received from an Edge VSAN to the Transit VSAN for outbound traffic and conversely translates the VSAN of a received frame from the Transit VSAN to the appropriate Edge VSAN; and (viii) terminates all control traffic including FSPF, zone server, and name server in the adjacent Edge VSAN. Each of the functions performed by the Border Switches is defined in more detail below.

[0028] Referring to Figure 2, an exemplary SAN according to the present invention is shown. The SAN 10 in this example includes three Edge VSANs (VSAN=1, VSAN=2, and VSAN=3), a Transit VSAN, four switches A, B, C, and D and three end devices, pWWN1, pWWN2 and pWWN3. Host pWWN1 is included in VSAN1. A second Host pWWN2 is included in VSAN2. A storage device pWWN3 is included in VSAN3. The Transit VSAN 12 is used to switch traffic between the disconnected Edge VSAN1, VSAN2, and VSAN3. The Transit VSAN 12, as is described below, enables the Hosts pWWN1 and pWWN2 to access the storage device pWWN3 and vice versa. In Fibre Channel, end devices such as Hosts and storage devices communicate using only Fibre Channel addresses. With the development of the VSAN, as described in detail in the above-identified parent application, a tag qualifies a Fibre Channel address so that it can be routed through the Switching Fabric, while being transparent to the end devices.

[0029] The Transit VSAN 12 switches traffic from the disconnected Edge VSAN pairs 1-3 and 2-3 and vice-versa. The link between the Switches B and C can be

according to various embodiments a standard FC link or remote (e.g., FCIP, FC over DWDM, etc) link. The transit VSAN 12 is similar to any other VSAN and can potentially have its own end devices attached to it. The Transit VSAN 12 does not care about the nature of the neighbor Edge VSANs that it is switching traffic either to or from. In other words, a transit VSAN can switch traffic to or from another transit VSAN.

[0030] Referring to Figure 3, another SAN 20 topology illustrating Border Switches is shown. Border Switches are switches that link between or are part of both an Edge VSAN(s) and a Transit VSAN(s). In this example, VSAN1 includes a Host (pWWN1, FC_ID = 3.1.1) and Switches S1 through S5. VSAN2 includes a storage device (pWWN2, FC_ID = 2.1.1) and Switches S6 through S10. Assuming that a Transit VSAN includes Switches S2, S3, S4, S5, S8 and S10, then S2 and S4 are considered Border Switches even though they are not necessarily at the border of VSAN1. Traffic coming from the Transit VSAN to VSAN 1 is by Switch S3 and traffic from Switch S1 and directed to the Transit VSAN is switched by S4. Similarly, Switches S8 and S10 are Border Switches for VSAN2. In yet another example, Switches B and C are defined as Border Switches in Figure 2 because they are capable of switching traffic between the Transit VSAN 12 and VSAN1, VSAN2 and VSAN3.

[0031] Referring to Figure 4A, a frame 30 having an Extended Inter-Switch Link (EISL) format for use typically on a link carrying multiple VSANs is shown. While the EISL format can also be used on a link carrying a single VSAN, another option on a link carrying only a single VSAN, is to configure the ingress port of the switch with the VSAN to be associated with frames received on that port. The frame 30 includes a Start of Frame (SOF) field 32, an EISL header field 34, an EISL payload field 36, an EISL Cyclical Redundancy Check (CRC) field 38, and a End of Frame (EOF) field 40. The SOF field 32 is the start of the frame delimiter. The EISL payload field 36 includes the header and payload used with a standard ISL frame. The EISL CRC field 306 is used to hold a CRC value specific to the EISL.

[0032] The EISL CRC value differs or is a modification of a standard CRC value calculated for a regular ISL frame due to the corresponding longer length of the EISL

frame 30 with the appended EISL header field 34. The EOF field 40 delineates the end of the frame 30.

[0033] Referring to Figure 4B, the EISL header field 34 is illustrated. This field includes a plurality of sub-fields including an EISL indicator field 42 for indicating the presence of an EISL header field 34, an EISL version field 44 that indicates the version of EISL header. In one embodiment, the version field 44 is at least two bits wide, allowing the indication of up to four versions of EISL frames. Additional bits may be used if more versions of EISL frames need to be indicated. The frame type field 46 is used to indicate the type of traffic to be carried by the frame. Traffic types may include for example Ethernet, Fibre Channel, or Infiniband. In one embodiment, the frame type field 46 is four-bits wide, allowing up to sixteen different traffic types to be transmitted across the Fabric. Moreover, multiple VSANs, each capable of transmitting different traffic types, may be interconnected through the identification of a traffic type as provided in this field. The Multi-Protocol Label Switching (MPLS) field 48 indicates if the frame is carrying MPLS information such as label stack 56, which is a common forwarding mechanism for both Fibre Channel and Ethernet. In one embodiment, the MPLS field 48 is one bit wide. When set, it indicates that the EISL header 34 includes an MPLS label stack 56. Otherwise it is reset.

[0034] Priority field 50 indicate the user priority of the EISL frame 30. Priority may be defined in a number of ways. As one example, the user priority may be a generic numeric priority, without a guaranteed level of service. For instance, higher values represent higher user priority while lower values may represent lower priority. Higher priorities receive available bandwidth first, regardless of how much total bandwidth is available. As another example, the user priority may indicate a quality of service (QoS) of the payload of the EISL frame. Generally, the width of the Priority field 50 depends on the priority type and/or the number of priority levels.

[0035] The VSAN identifier field 52 or “tag” is used to identify the frame 30 as belonging to a particular VSAN. More particularly, the VSAN identifier field 52 identifies the payload of the EISL frame 30 as belonging to a particular VSAN. In accordance with one embodiment, the VSAN identifier field 412 is a twelve-bit wide field. The format of the identifier may be identical to or similar to VLAN identifiers

as well as similar to addresses employed in certain standard protocols such as Ethernet.

[0036] In some SANs, there may be topology as well as routing problems that could cause a frame to traverse a loop within the network. Such a loop will consume bandwidth unnecessarily. In order to address this problem, a Time To Live (TTL) field 54 may be used to indicate a TTL value specifying the number of remaining hops that can be traversed before the frame is dropped. The TTL value inserted into field 54 is initialized by the network device (e.g., a Switch) that generates the EISL frame 30. A TTL default value is initially set to an arbitrary number, for example sixteen. With each hop, subsequent network devices (e.g., Switches) receiving the EISL frame decrement the TTL value by one. A TTL value of one indicates to the receiving network device (e.g., Switch) that the EISL frame should be dropped. When the EISL frame is dropped, an error message may be sent to the intended recipient of the frame as well as to the sender of the frame. Similarly, a TTL value of 0 may indicate that the TTL field 54 should be ignored, allowing the EISL frame to be forwarded by the switch.

[0037] On a link carrying multiple VSANs, Switches communicate using frames 30. Each frame 30 also includes, in addition to the above, the Fibre Channel addresses (FC_IDs) of the source and destination end devices. The VSAN ID 52 qualifies a particular frame 30 as belonging to a particular VSAN, transparent to the end devices. For more information on the switching of frames 30 within a VSAN, see the aforementioned parent application.

[0038] Referring to Figure 5, a VSAN 60 including “Inter-VSAN zones” according to the present invention is shown. An Inter-VSAN zone is defined as a zone that has members from multiple VSANs. In this example, Inter-VSAN zone X includes Host pWWN1 (FC_ID 1.1.1) in VSAN 1 and storage device pWWN3 (FC-ID 8.1.1) in VSAN 3. Inter-VSAN zone Y includes Host pWWN2 (FC_ID 3.1.1) in VSAN 2 and the same storage device pWWN3 (FC-ID 8.1.1) in VSAN 3. Inter-VSAN zones X and Y have thus been created so that Hosts pWWN1 and pWWN2 can both communicate with storage device pWWN3 respectively.

[0039] Using the inter-VSAN zone, Border Switches decide: (i) the content of the

name server database that is exported into the Transit VSAN from the adjacent Edge VSAN and vice versa; (ii) the set of FSPF domains to export in Link State Update (LSU) messages; (iii) the set of addresses to switch from adjacent Edge VSANs to Transit VSANs and vice versa; (iv) the set of adjacent Edge VSANs to which SW_RSCNs received from a Transit VSAN are propagated; (v) the set of SW_RSCNs received from adjacent Edge VSANs to propagate into the Transit VSAN. In other words, the Inter-VSAN zone is the point from which the import and export data and control traffic occurs. Since zone configuration is a well known concept, the configuration of Inter-VSAN routing via Inter-VSAN zones simplifies management and administration in the Fabric.

[0040] Similar to a regular zone in a VSAN, Inter-VSAN zones are contained within a zoneset and there is an active zoneset. The Border Switches thus determine the import and export traffic from the Inter-VSAN zones in the active zoneset.

[0041] In a typical SAN, a number of protocols are implemented when a link is established between two Switches. These include the Port Manager; Domain Manager; Zone Server; Fabric Shortest Path First (FSPF); Name_Server; and Switch Register State Change Notification (SW_RSCN) protocols, as described in the above-referenced NCITS documents. The aforementioned protocols have been modified for Inter-VSAN routing as contemplated with the present invention as described below.

[0042] **Port Manager Protocol:** The Port manager protocol negotiates parameters between two Switches and determines if the link is an Inter-Switch Link (ISL) or an Extended Inter-Switch Link (EISL). If the link is EISL, then it means that the link is capable of carrying VSAN frames. Otherwise, the Port Manager protocol operates the same with VSANs as with a regular SAN.

[0043] **Domain Manager Protocol:** The Domain Manager Protocol is responsible for the allocation of Domain_IDs and/or Fibre Channel addresses (FC_ID) for each Switch, Host and storage device in the SAN. As noted above, a FC_ID includes three components, a Domain_ID, an Area_ID, and a Port_ID. During initialization, a Principal Switch is selected for the SAN. The Principal Switch is responsible for assigning a Domain_ID for each Switch. Each Switch is then responsible for selecting the Area_ID and Port_ID for each end device connected to that Switch.

[0044] According to the present invention, the domain number space must be unique across the VSANs that are going to communicate with one another. There are a number of ways in which this uniqueness can be maintained, including: (i) administratively partitioning the domain number space across the VSANs; or (ii) associate a set of Domain_IDs to be used only for Inter-VSAN routing. For example, Domain_IDs between 200-239 (or any other arbitrary range) can be dedicated for VSAN routing. A Switch that needed to communicate across VSANs could administratively be configured to request Domain_IDs in the dedicated number space range.

[0045] **Zone Server Protocol:** In a standard SAN, the Zone Server Protocol is responsible for creating and maintaining a database within each Switch that defines each zone in the SAN. The zones in the table are defined in terms of the Host(s) and storage device(s) in each zone. A Host and/or storage device can belong to multiple zones. To ensure a consistent zoning database across a SAN, when an ISL link comes up between two switches, the zone database is exchanged between the switches to ensure consistency and to prevent conflicting zone definitions. If there are no conflicts, the zone databases are merged. If there are conflicts, the link is isolated and no data traffic can flow through that link until the conflicts are resolved. As part of the zone server protocol, whenever a zone is added, deleted or modified, the changes are propagated throughout the Fabric.

[0046] To support Inter-VSAN routing, the Zone Server Protocol is modified to accommodate Inter-VSAN zones having members in different VSANs. Further, existing mechanisms or new mechanisms can be devised to ensure consistency of inter-VSAN zones. Since VSANS are terminated at the border switches, intra-domain zones are not propagated to the Transit VSAN.

[0047] **Fabric Shortest Path First (FSPF) Protocol:** The FSPF is a link state path selection protocol. FSPF keeps track of the state of the links on all the Switches in the Fabric and associates the cost with each link. The protocol computes paths from each Switch to all the other Switches in the Fabric by adding the cost of all the links traversed by the path, and choosing or selecting the path that minimizes the cost. The collection of the link state records (LSR's) (including the cost) of all the Switches in

the Fabric constitutes the topology database of the Fabric, called the Link State Database.

[0048] The FSPF protocol has four major components, including: (i) a “Hello” protocol used to establish connectivity between neighbor Switches, to establish the identity of neighbor Switches, and to exchange FSPF parameters and capabilities between the neighbor Switches; (ii) a replicated fabric topology or Link State Database, with protocols and mechanisms to keep the databases synchronized across the Fabric; (iii) a path computation algorithm; and (iv) a routing table update.

[0049] The Link State Database synchronization in turn consists of two major components, an initial database synchronization and an update mechanism. The initial database synchronization is used when a Switch is initialized or when an inter-Switch Link (ISL) comes up. The update mechanism is used when either (i) there is a link state change, for example when an ISL goes up or down; or (ii) on a periodic basis, to prevent Switches from deleting topology information from the database.

[0050] With the FSPF protocol, the term “path selection” indicates the lowest cost or “best” path between a source and destination in the Fabric. The term “routing” indicates the actual forwarding of frames to a specific destination. FSPF performs hop-by-hop routing, which means that a Switch in the Fabric only needs to know the next hop on the best path to the destination. The replicated topology database insures that every Switch in the Fabric has the same definition of the Fabric, and therefore, all the Switches will make consistent routing decisions. Typically, a Switch needs to know, for each destination domain in the Fabric, which path should be used to route a frame to a domain. A routing table entry therefore requires at a minimum a destination Domain_ID and an E_Port to which frames are forwarded.

[0051] Since the FSPF protocol is contained within a VSAN, in order to support routing for domains that are in other VSANs, the following modifications are implemented:

[0052] For each Domain_ID of a device that is part of an inter-VSAN zone, a border switch considers the Domain_ID for announcement into the adjacent VSANs. An adjacent VSAN may be a transit VSAN or an edge VSAN. A thus selected Domain

ID is announced into an adjacent VSAN if either that VSAN is a transit VSAN or there is a device in the VSAN that is part of the same inter-VSAN zone.

[0053] In addition to this, a border switch rewrites the VSAN of a frame that is being routed across VSANs to that of the adjacent VSAN. So, for frames being routed from an edge VSAN to a transit VSAN, a border switch rewrites the VSAN from the edge VSAN to the transit VSAN.

[0054] Referring to Figure 6, an exemplary SAN 70 useful for illustrating the modifications to the FSPF protocol is shown. The SAN 70 includes VSAN1, VSAN2, a transit VSAN and an inter-fabric zone Z with members pWWN1, pWWN2 and pWWN4. With a standard SAN, Switch C4 would announce the LSRs of its Link State Database to itself, Switch C3, Switch E3, and Switch E4, where each of the LSRs would contain information about all the links connected to the Switch C4. With Inter-VSAN routing, Switch C4 announces only a route for Switch E4 with a link cost. Similarly, Switch E3 would announce only a route for E4. Similarly, switches C1 and C2 each announce a route for E2 and E1. On receiving a route announcement from Switches C3 and C4, Switch C1 and C2 announce the route for Switch E4 obtained from the Transit-VSAN into VSAN 1. Likewise, Switches C3 and C4 announce the route for Switches E1 and E2 into VSAN 2. From the forwarding table perspective, Switches C1 and C2 have to change the VSAN for frames destined to E4 from VSAN1 to the Transit VSAN. In the opposite direction, they will change the VSAN for frames received for Switches E1 and E2 from the Transit VSAN to VSAN1. Similarly, Switches C3 and C4 will change the VSAN for frames destined to E1 and E2 from VSAN 2 to the Transit VSAN and change the VSAN for the frames received for Switch E4, from the Transit VSAN to VSAN 2.

[0055] **Name Server Protocol:** With the Name Server Protocol, each Switch exchanges information regarding its locally attached end devices with the other Switches in the SAN. The information that is exchanged for each end device includes the world wide name (pWWN), the Fibre Channel address (FC_ID), the type of protocol (SCSI, IP, etc.) the end device supports, and if the end device is an initiator (i.e., a Host) or a target (i.e., a storage device).

[0056] To support Inter-VSAN routing, the changes to the name server on a Border Switch are: (i) build the list of name server entries to be exported to a Transit VSAN or Edge VSAN. This list is created from the defined Inter-VSAN zones, and (ii) terminate and proxy name server queries for the domains associated with the remote domain and respond on behalf of the name server of the queried Switch.

[0057] Referring again to Figure 6, the modification rules of the Name Server Protocol can be illustrated by way of examples.

- 1) Switches C1 and C2 add to the Name Server database in the Transit VSAN the devices pWWN1 and pWWN2. Similarly, Switches C3 and C4 add pWWN4 to the Name Server database in the transit VSAN. When C4 announces the route to domain E3 in the transit VSAN and C1 and C2 receive this announcement, each of them send a name server query for information about all attached end devices to E3 to the name server of E3 as specified by the FC standards. Depending on the route taken by the query, either C3 or C4 intercept this query and respond on behalf of E3 with information only about pWWN4. Likewise, when C1 and C2 announce the route to domains E1 and E2, it is received by C3 and C4. C3 and C4 then send name server queries to E1 and E2 which are intercepted by C1 or C2 which proxy and respond on behalf of the name server for the announced domains. C1 and C2 respond only with information about pWWN1 for queries to E1 and with pWWN2 for queries to E2.
- 2) A similar process occurs in VSAN 1 when the route to domain E4 is announced by C1 and C2. Each switch in VSAN 1, in this case, E1 and E2, send name server queries for all attached end devices to E4 to the name server of E4. Depending on the route taken by the query, either C1 or C2 intercept the query and respond with information only about pWWN4.
- 3) A similar process as above occurs in VSAN 2 when C3 and C4 announce routes to domains E1 and E2.
- 4) If a device is added to an already announced domain or if the state of the device announced changes, notification of a change is conveyed via

SW_RSCN. When a switch receives an SW_RSCN, it generates a name server query to the switch to which the end device identified in the SW_RSCN is attached. This query is also intercepted and responded to by the border switches.

Register State Change Notification (SW_RCSN) Protocol

[0058] SW_RSCNs are exchanged between switches whenever the name server database changes. The change is usually due to a change in the state of a locally connected port, a locally connected switch or to the zone server database. SW_RSCNs are originated by the Switch detecting the change and sent to every other Switch in the VSAN. The SW_RSCN contains information about the affected end devices or Domain_ID (Domain_ID is used when changes affect the entire switch as specified in the FC-MI Technical Report, incorporated by reference herein for all purposes. This information includes the port WWN of the end device and its FC_ID. An SW_RSCN can contain notification about multiple end devices.

[0059] With Inter-VSAN routing, changes in one VSAN must be propagated to other VSANs if the change affects a device in an Inter-VSAN zone. Consider the topology in Figure 6. If switch E4 loses the connection to pWWN4, it transmits an SW_RSCN to C3 and C4. Since pWWN4 is in a Inter-VSAN zone and C3 and C4 are Border Switches, they both will propagate the SW_RSCN to the Transit VSAN. This is received by both C1 and C2 which in turn propagate the SW_RSCN in VSAN 1. Thus, duplicate SW_RSCNs are generated which is inefficient.

[0060] The replication of SW_RSCN frames can be prevented by selecting a single Switch to be responsible for distributing SW_RSCN from one VSAN to another. The selection of such a switch can be done in multiple ways, for example, either: (i) statically configured by a system administrator; (ii) selecting the principal Switch in a VSAN to be designated as the Switch for distributing SW_RSCNs; (iii) the border switches can then announce the SW_RSCN from an adjacent VSAN to only the principal switch which then distributes the SW_RSCN within its VSAN or (iv) one of the Border Switches can be selected based on some scheme such as shortest path to the VSAN for which it is advertising, the switch with the highest “switch WWN” etc. Regardless of the responsible Switch, each switch receiving a SW_RSCN notifies

registered end hosts of the change as per normal RSCN rules.

EXAMPLES

[0061] Figure 7 presents a scenario where two devices, pWWN1 and pWWN2, need to communicate with a storage device pWWN3. The communication between pWWN2 and pWWN3 is similar to the other scenarios that have been discussed previously and therefore will not be discussed in detail herein. However, consider the case of pWWN1 communicating with pWWN3 where there is not a direct connection between the Border Switches of VSAN 2 and VSAN 3. Figure 7 illustrates two subcases of this topology, when the two transit VSANs are the same and when they are not the same.

[0062] When the transits VSANs are the same, in order to have pWWN1 communicate with pWWN3, an Inter-VSAN zone is defined containing the two devices. There must be one or more links connecting switches S2 and S3 that carry the Transit VSAN traffic. If the link between the switches is not direct, all the switches in the path between S2 and S3 must carry the transit VSAN.

[0063] When the transit VSANs are not the same, the Inter-VSAN zones must be defined such that frames can be switched from one Transit VSAN to the other. In figure 7, a route to the switch associated with PWWN1 is announced in the transit VSAN by S1. S2 on hearing this route announcement as specified above, announces it in VSAN 1. In VSAN 1, when S3 hears this announcement, it propagates the route announcement to Transit VSAN V2, where S4 then announces in VSAN 3. Now a path has been established between the switches connected to end devices PWWN1 and PWWN3 and so communication can take place. Along with route announcements, name server query proxying and SW_RSCN propagation as specified above also occurs.

[0064] Referring to Figure 8, a block diagram of a Switch that can be used for Inter-VSAN routing in a SAN according to the present invention is shown. The Switch 80 includes a data plane 82 and a control plane 84. In the data plane 82, the Switch includes switching logic 86 connected between two sets of ports 88a and 88b. The switching logic is configured to route or internally switch traffic received on one port 88a to another port 88b and vice versa. The control plane includes a state machine 90

for implementing the Port Manager; Domain Manager; Zone Server; Fabric Shortest Path First (FSPF); Name_Server; and Register State Change Notification (RSCN) protocols and their modification and changes as described herein. In alternative embodiments, the protocols can be implemented in hardware, programmable logic, in software running on a micro-controller, or a combination thereof.

[0065] The embodiments of the present invention described above are to be considered as illustrative and not restrictive. The various change commands described herein are only exemplary and any other types of commands may be used. The invention is not to be limited to the details given herein, but may be modified within the scope and equivalents of the appended claims.